Avaliação de Desempenho
Tópicos Especiais em Arquitetura de Computadores
Terceiro Trabalho Prático (v2)
Análise de Grupos

Observações gerais sobre este trabalho

1. Este é um exercício de programação de análises de dados sobre bases de dados;
2. Você poderá programar em qualquer linguagem ou ambiente que você queira (Python, Shell script, R). Entretanto, a linguagem Python já tem muitas bibliotecas prontas (sci-kit learn, numpy e pandas, por exemplo);
3. Se for aluno de graduação, poderá fazer este trabalho individual, ou em duplas; se for aluno de pós-graduação terá que fazer o trabalho individual;
4. Há diversos materiais disponíveis na Web sobre análise de grupos, os quais podem ser consultados sem qualquer problema;
5. O livro texto dessa disciplina tem muitas dicas importantes e exemplos;
6. Você pode se basear no Jupyter Notebook disponibilizado nos materiais de estudos;
7. Considere a base de dados **Wine**;
8. Todos os itens abaixo valem **1.0** ponto, exceto o sexto e o oitavo itens que valem 2.0 pontos.

Você deve escrever um **Notebook no Google Colaboratory** que resolva os problemas descritos abaixo:

1) O dataset Wine possui 14 colunas. A primeira coluna refere-se a classe dos vinhos. Gere um novo dataframe **sem** a primeira coluna. Em seguida, faça o pré-processamento dos dados de tal forma que **todos** os **treze** atributos estejam normalizados no intervalo [0,1]. Imprima as primeiras 10 linhas antes e depois da normalização;

2) Apesar do K-Means poder ser utilizado para múltiplos atributos, para facilitar a aplicação e melhoria da visualização dos dados, aplique a técnica ***Principal Component Analysis*** (PCA) para reduzir os dados de 13 para 2 atributos;

3) A medida de similaridade para os objetos será feita, para efeitos comparativos, a partir de duas medidas: a distância Euclidiana e a correlação de Pearson (por favor, pesquise sobre esse assunto). Defina funções específicas para cada caso;

4) Serão aplicados dois algoritmos de agrupamento: k-médias e DBSCAN. Estes dois algoritmos serão aplicados considerando as duas medidas de similaridade. Para o algoritmo k-médias, justifique o valor de k adotado. Sugestão, pesquise e aplique o método "cotovelo" para identificar o valor de k. Para o algoritmo DBSCAN, justifique os valores dos parâmetros: raio de vizinhança e quantidade mínima de pontos.

5) Somente para o algoritmo k-médias, imprima todos os centróides.

6) Imprima uma tabela com todos os atributos ordenados por cada grupo, considerando as quatro possibilidades, isto é, k-médias com distância Euclidiana, k-médias com correlação de Pearson, DBSCAN com distância Euclidiana e DBSCAN com correlação de Pearson.

7) Escolha e gere alguma forma de visualização dos grupos (as quatro possibilidades). Pode ser gráfico (chamados de protótipos), grafo, dendograma (também chamados de árvores), ou qualquer outro método que você julgar procedente;

8) Avalie os quatro agrupamentos obtidos por duas medidas internas: índice de Dunn e índice Silhueta.